Davide Belli
11887532
davidebelli95@gmail.com

Homework Assignment 5
Machine Learning 2, 17/18

2018-05-05

In this Homework, I discussed possible interpretations with Gabriele Cesa.

1 Problem 1

1.1 Question 1

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \pi_j} \mathbb{E}_{posterior}[\log p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] + \lambda(\sum_k \pi_k - 1) \\
0 &= \frac{\partial}{\partial \pi_j} \sum_n \sum_k \gamma(z_{nk}) \{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} + \lambda(\sum_k \pi_k - 1) \\
0 &= \sum_n \frac{\gamma(z_{nj})}{\pi_j} + \lambda \\
-\sum_n \gamma(z_{nj}) &= \lambda \pi_j \\
-\sum_j N_j &= \lambda \sum_j \pi_j \\
-N &= \lambda
\end{aligned}
$$

$$
\begin{aligned}
-N_j &= \lambda \pi_j \\
-N_j &= -N \pi_j \\
\pi_j &= \frac{N_j}{N}
\end{aligned}
$$

$$
0 = \frac{\partial}{\partial \boldsymbol{\mu}_j} \mathbb{E}_{posterior}[\log p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})]
$$

$$
0 = \frac{\partial}{\partial \boldsymbol{\mu}_j} \sum_n \sum_k \gamma(z_{nk}) \{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}
$$

$$
0 = -\frac{1}{2} \sum_n \gamma(z_{nj})(-2\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_j)
$$

$$
\begin{aligned}
\sum_n \gamma(z_{nj}) \boldsymbol{\Sigma}_j^{-1} \boldsymbol{x}_n &= \sum_n \gamma(z_{nj}) \boldsymbol{\Sigma}_j^{-1} \boldsymbol{\mu}_j \\
\sum_n \gamma(z_{nj}) \boldsymbol{x}_n &= \boldsymbol{\mu}_j \sum_n \gamma(z_{nj}) \\
\boldsymbol{\mu}_j &= \frac{\sum_n \gamma(z_{nj}) \boldsymbol{x}_n}{N_j}
\end{aligned}
$$

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\Sigma}_j} \mathbb{E}_{posterior}[\log p(\boldsymbol{X}, \boldsymbol{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \\
0 &= \frac{\partial}{\partial \boldsymbol{\Sigma}_j} \sum_n \sum_k \gamma(z_{nk}) \{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \\
0 &= \frac{1}{2} \sum_n \gamma(z_{nj})(\boldsymbol{\Sigma}_j^{-1} - \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_j)(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}) \\
\sum_n \gamma(z_{nj}) \boldsymbol{\Sigma}_j^{-1} &= \sum_n \gamma(z_{nj}) \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_j)(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} \\
\sum_n \gamma(z_{nj}) &= \sum_n \gamma(z_{nj})(\boldsymbol{x}_n - \boldsymbol{\mu}_j)(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}
\end{aligned}
$$

$$\mathbf{\Sigma}_j^{-1} = \frac{N_j}{\sum_n \gamma(z_{nj})(\boldsymbol{x}_n - \boldsymbol{\mu}_j)(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T}$$
$$\mathbf{\Sigma}_j = \frac{\sum_n \gamma(z_{nj})(\boldsymbol{x}_n - \boldsymbol{\mu}_j)(\boldsymbol{x}_n - \boldsymbol{\mu}_j)^T}{N_j}$$

1.2 Question 2

When the covariance matrices have a common value $\mathbf{\Sigma}$, the update rules for $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ do not change, since they do not depend on the covariance. On the other hand, the update rule for $\mathbf{\Sigma}$ becomes:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \mathbf{\Sigma}} \mathbb{E}_{posterior}[\log p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\mu}, \mathbf{\Sigma}, \boldsymbol{\pi})] \\
0 &= \frac{\partial}{\partial \mathbf{\Sigma}} \sum_n \sum_k \gamma(z_{nk}) \{\log \pi_k + \log \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \mathbf{\Sigma})\} \\
0 &= \frac{1}{2} \sum_n \sum_k \gamma(z_{nk})(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}) \\
0 &= \frac{1}{2} \sum_n \sum_k \gamma(z_{nk})(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1}) \\
\sum_n \sum_k \gamma(z_{nk}) \mathbf{\Sigma}^{-1} &= \sum_n \sum_k \gamma(z_{nk}) \mathbf{\Sigma}^{-1}(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1} \\
\sum_n \sum_k \gamma(z_{nk}) &= \sum_n \sum_k \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}^{-1} \\
\mathbf{\Sigma}^{-1} &= \frac{N}{\sum_n \sum_k \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T} \\
\mathbf{\Sigma} &= \frac{\sum_n \sum_k \gamma(z_{nk})(\boldsymbol{x}_n - \boldsymbol{\mu}_k)(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T}{N}
\end{aligned}
$$

2 Problem 2

Let's decompose the log-posterior in:

$$
\begin{aligned}
\log p(\boldsymbol{\theta}|\boldsymbol{X}) &= \log p(\boldsymbol{X}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{X}) = \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + KL(q||p) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{X})
\end{aligned}
$$

In the EM algorithm, in order to maximize directly the posterior, we instead maximize the ELBO, defined as:

$$ELBO = \mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{X})$$

Since in th E-step we maximize it with respect to q, we notice that we only care about the first term, since the logprior and log evidence are constant with respect to q. It is then evident as this E-step is the same as in the likelihood maximization case, where the ELBO is:

$$ELBO_{likelihood} = \mathcal{L}(q, \boldsymbol{\theta})$$

On the other hand, in the M-step we maximize the ELBO with respect to the parameters $\boldsymbol{\theta}$, thus we need to include the logprior in the mazimization quantity defined as:

$$
\begin{aligned}
argmax_{\boldsymbol{\theta}} ELBO &= argmax_{\boldsymbol{\theta}} \big(\mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{X})\big) = \\
&= argmax_{\boldsymbol{\theta}} \big(\mathcal{L}(q, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))\big) =
\end{aligned}
$$

$$= argmax_{\boldsymbol{\theta}}\big(\mathbb{E}_{posterior}[\log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})] + H(p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old})) + \log p(\boldsymbol{\theta}))\big) =$$
$$= argmax_{\boldsymbol{\theta}}\big(\mathbb{E}_{posterior}[\log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})] + \log p(\boldsymbol{\theta}))\big) =$$
$$= argmax_{\boldsymbol{\theta}}\big(\sum_z p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))\big) =$$

3 Problem 3

Using the result from the previous task, the M-step is:

$$\begin{aligned}
argmax_{\boldsymbol{\theta}} ELBO &= argmax_{\boldsymbol{\theta}}\big(\mathcal{L}(q,\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) - \log p(\boldsymbol{X})\big) = \\
&= argmax_{\boldsymbol{\theta}}\big(\sum_z p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{old}) \log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}))\big) = \\
&= argmax_{\boldsymbol{\mu},\boldsymbol{\pi}}\big(\sum_z p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\mu}^{old},\boldsymbol{\pi}^{old}) \log p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\mu},\boldsymbol{\pi}) + \log p(\boldsymbol{\mu}) + \log p(\boldsymbol{\pi}))\big) \\
&= argmax_{\boldsymbol{\mu},\boldsymbol{\pi}}\Bigg(\sum_n\sum_k \gamma(z_{nk})\Big(\log \pi_k + \sum_i^D [x_{ni}\log \mu_{ki} + (1-x_{ni})\log(1-\mu_{ki})]\Big)\Bigg) \\
&+ \sum_k\sum_i^D \Big((a_k-1)\log \mu_k + (b_k-1)\log(1-\mu_{ki}) - \log B(a_k,b_k)\Big) \\
&+ \sum_k \Big((\alpha_k-1)\log \pi_k - \log B(\alpha_k)\Big) + \lambda(\sum_k \pi_k - 1)
\end{aligned}$$

For $\boldsymbol{\mu}$:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \mu_{ki}}\Bigg(\sum_n\sum_k \gamma(z_{nk})\Big(\log \pi_k + \sum_i^D [x_{ni}\log \mu_{ki} + (1-x_{ni})\log(1-\mu_{ki})]\Big)\Bigg) \\
&+ \sum_k\sum_i^D \Big((a_k-1)\log \mu_{ki} + (b_k-1)\log(1-\mu_{ki}) - \log B(a_k,b_k)\Big) \\
&+ \sum_k \Big((\alpha_k-1)\log \pi_k - \log B(\alpha_k)\Big) + \lambda(\sum_k \pi_k - 1) \\
0 &= \sum_n \gamma(z_{nk})\Big(\sum_i^D \frac{x_{ni}}{\mu_{ki}} - \frac{1-x_{ni}}{1-\mu_{ki}}\Big) + \frac{a_k-1}{\mu_{ki}} - \frac{b_k-1}{1-\mu_{ki}} \\
\frac{\Big((a_k-1) + \sum_n \gamma(z_{nk})x_{ni}\Big)}{\mu_{ki}} &= \frac{\Big((b_k-1) + \sum_n \gamma(z_{nk})(1-x_{ni})\Big)}{1-\mu_{ki}} \\
\Big((a_k-1) + \sum_n \gamma(z_{nk})x_{ni}\Big) &= \mu_{ki}\Big((a_k-1) + \sum_n \gamma(z_{nk})x_{ni} + (b_k-1) + \sum_n \gamma(z_{nk})(1-x_{ni})\Big) \\
\Big((a_k-1) + \sum_n \gamma(z_{nk})x_{ni}\Big) &= \mu_{ki}\Big(a_k + b_k - 2 + \sum_n \gamma(z_{nk})\Big) \\
\mu_{ki} &= \frac{a_k - 1 + \sum_n \gamma(z_{nk})x_{ni}}{a_k + b_k - 2 + N_k}
\end{aligned}$$

While, for $\boldsymbol{\pi}$:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \pi_k}\Bigg(\sum_n\sum_k \gamma(z_{nk})\Big(\log \pi_k + \sum_i^D [x_{ni}\log \mu_{ki} + (1-x_{ni})\log(1-\mu_{ki})]\Big)\Bigg) \\
&+ \sum_k\sum_i^D \Big((a_k-1)\log \mu_{ki} + (b_k-1)\log(1-\mu_{ki}) - \log B(a_k,b_k)\Big) \\
&+ \sum_k \Big((\alpha_k-1)\log \pi_k - \log B(\alpha_k)\Big) + \lambda(\sum_k \pi_k - 1)
\end{aligned}$$

$$
\begin{aligned}
0 &= \frac{\sum_n \gamma(z_{nk}) + \alpha_k - 1}{\pi_k} + \lambda \\
0 &= \sum_n \gamma(z_{nk}) + \alpha_k - 1 + \lambda \pi_k \\
\lambda \sum_k \pi_k &= -(\sum_k \sum_n \gamma(z_{nk}) + \sum_k \alpha_k - \sum_k 1) \\
\lambda &= -(N + \sum_k \alpha_k - K)
\end{aligned}
$$

$$
\begin{aligned}
0 &= \sum_n \gamma(z_{nk}) + \alpha_k - 1 + \lambda \pi_k \\
(N + \sum_k \alpha_k - K)\pi_k &= \sum_n \gamma(z_{nk}) + \alpha_k - 1 \\
\pi_k &= \frac{\sum_n \gamma(z_{nk}) + \alpha_k - 1}{N + \sum_k \alpha_k - K}
\end{aligned}
$$